Stability analysis of a Basic Collaboration System via fluid limits

Rosario Delgado[1⋆] and Evsey Morozov[2 ⋆⋆]

[1] Departament de Matemàtiques. Universitat Autònoma de Barcelona.
Edifici C- Campus de la UAB. Av. de l'Eix Central s/n.
08193 Bellaterra (Cerdanyola del Vallès)- Barcelona, Spain
delgado@mat.uab.cat
[2] Institute of Applied Mathematical Research.
Russian Academy of Sciences and Petrozavodsk State University, Russia
emorozov@karelia.ru

Abstract. In this work, the fluid limit approach methodology is applied to find a sufficient and necessary stability condition for the Basic Collaboration (BC) system with feedback allowed, which is a generalization of the so-called W-model. In this queueing system, some customer classes need cooperation of a subset of (non-overlapping) servers. We assume that each customer class arrives to the system following a renewal input with general i.i.d. inter-arrival times, and general i.i.d. service times are also assumed. Priority is given to customer classes that can not be served by a single server but need a cooperation.

Keywords: stability; fluid limit approach; Skorokhod problem; workload; BC system; W-model

1 Introduction

In this paper, we study a generalization of the so-called queueing W-model which, in the simplest setting, consists of two single-server stations, 1, 2, and three infinite-capacity buffers, 1, 2, 3, with independent renewal inputs of class-k customers, respectively, $k = 1, 2, 3$. Server i processes class-i costumers, $i = 1, 2$, but both servers are required to process class-3 customers which have preemptive-resume priority. (For more detailed description see [9, 14].)

We generalize the W-model, which is in turn a particular case of the so-called *sparsely connected* model [14]. More exactly, we consider a Basic Collaboration (BC) system with J infinite buffer servers and $K \geq J$ customer classes. Each customer class needs cooperation of a subset of (non-overlapping) servers (it is called *concurrent service*). At the same time, there may be customer classes that

⋆ Supported by Ministerio de Economía y Competitividad, Gobierno de España, project ref. MTM2015 67802-P

⋆⋆ Supported by Russian Foundation for Basic Research, projects 15-07-02341, 15-07-02354, 15-07-02360 and by the Program of strategy development of Petrozavodsk State University

only need a server to be served. Overlapping customer classes on a server can only occur between a class that needs its cooperation with another server(s), and a class that only needs it to be served, without cooperation. In this setting, to keep work-conserving service discipline, we assume the mentioned priority of the customer requiring cooperation. We assume i.i.d. general inter-arrival and i.i.d. service times. Such a system is also called joint service model [10], or concurrent server release [2]. Queueing systems with concurrent service have been considered in a number of works, and pioneering ones are [12, 2, 17, 8]. For the buffer-less (loss) concurrent service systems, the performance analysis has been developed in a number of works [1, 11, 17, 16, 15]. However, analysis of the buffered concurrent service system is much more challenging.

A comprehensive study of the concurrent service system has been developed in [12], where the author used the matrix analytic method to deduce a stability condition. However, this condition requires to solve a matrix equation of a large dimension, and moreover, the corresponding matrices are not explicitly defined. The authors of [13] study a multi-server system in which each customer requires a random number of servers simultaneously and a random but identical service time at all occupied servers, which describes the dynamics of modern high performance clusters. They assume exponential distributions and an arbitrary number of servers. In [13], a modification of the matrix-analytic method is developed to obtain stability criterion of the simultaneous service model in an explicit form. (Also see [13] for a broad bibliographic review on the subject including previous references.) Note that the paper [14] considers various sparsely connected models assuming saturated regime, while, in the present research, we are seeking for the stability conditions.

Motivation. BC systems model real situations in which different agents are able to work together to solve complex problems. Consider the following scenario introduced in [18]. A user wishes to determine the best package price for a ski trip given the following criteria: a resort in the Alps, for a week in February, with slope-side lodging, and the lowest price for all expenses. To solve this problem, an agent obtains a list of appropriate ski resorts from a database before spawning other agents to query travel databases, possibly in different formats, for package prices at those resorts in February. Agents can perform this task more efficiently when they can correlate their results and adjust their computations based on the outcome of a collaboration. Suppose the agents visit local travel agencies and then share their intermediate results and collaborate before migrating to another travel agency. If an agent determines that a particular resort does not have any available lodging meeting the user's criteria, the agents may determine to drop queries about trips to that destination. As more information is gathered, agents may also make other decisions. As this example demonstrates, agents can perform complex distributed computations more effectively if they based on the combined results. To do it, they can divide a complex task into smaller pieces and delegate them to agents that migrate throughout the network to accomplish them. These agents perform computations, synchronously share results,

and collaboratively determine any changes to future actions, giving service to the user.

Another example are medical centers and hospitals, in which different types of patients have different requirements concerning technical equipment, facilities, doctors and nurses, which can be considered as the *servers*.

We give a brief summary of the research. The main contribution of this work is that, in contrast to previous works on W-models and concurrent service systems in general, we obtain stability condition, following fluid stability analysis developed in [3]. Indeed, our model is more general than the *Generalized Jackson network* in [3] (Section 5) , in which there is only one class of customers served at each single-server station. Instead, in our model each server can serve more than one class: at most one customer class requires cooperation with other servers (*multiserver* customers), but no limit on the number of customer classes that do not need cooperation (*single-server* customers). Note that multi-class customers but in a single-station network have been considered in [3] (Section 6), as well. In Theorem 1 we first establish the stability of the fluid limit model associated to the BC system under sufficient condition. The fluid limit model, which allows to transform the initial stochastic problem into a (related) deterministic one, is introduced in Proposition 1. The stability of the fluid limit model means that the fluid limit of the queue-size process reaches zero in a finite time interval and stays there. Then, using stability of the fluid limit model and Theorem 4.2 [3], we deduce positive Harris recurrence of the basic Markov process describing the network. Similarly to [3], functional laws of large numbers for the renewal processes or, in other words, the hydrodynamic scaling by the increasing value of the initial state, are used to obtain the stability of the fluid limit model via the solution of a Skorokhod problem. At that, the choice of an appropriate Lyapunov function is the key point of analysis. By the same approach, we show that if the necessary condition is violated, then the fluid limit model is *weakly unstable*. It means that, if the process starts at zero, then there exists a time at which the fluid limit of the queue-size process becomes positive. As a result, by Theorem 3.2 [4], the queueing network is *unstable*: the queue size grows infinitely with probability (w.p.) 1 as time increases.

The paper is organized as follows. In section 2, we give notation and describe the BC in more detail, introducing the associated queueing network equations. Section 3 contains fluid stability analysis, at that, in section 3.1, the fluid limit model is constructed, and the proof of stability condition is given in section 3.2 (Theorem 2).

2 Notation and description of the BC

We first give basic notation. Vector are column vectors and (in)equalities are interpreted component-wise. v^T denotes the transpose of a vector (or a matrix). For any integer $d \geq 1$, let $\mathbb{R}_+^d = \{\, v \in \mathbb{R}^d \,:\, v \geq 0 \,\}$, $\mathbb{Z}_+^d = \{\, v = (v_1, \ldots, v_d)^{\mathbf{T}} \in \mathbb{R}^d \,:\, v_i \in \mathbb{Z}_+ \,\}$. For a vector $v = (v_1, \ldots, v_d)^{\mathbf{T}} \in \mathbb{R}^d$, let $|v| = \sum_{i=1}^d |v_i|$. We

denote $diag(v)$ the diagonal matrix with diagonal entries being the components of vector v, and I is the d-dimensional identity matrix. We say that a sequence of vectors $\{v^n\}_{n\geq 1}$ converges to a vector v as $n\to\infty$ if $|v^n - v|\to 0$, and denote it as $\lim\limits_{n\to\infty} v^n = v$. (This convergence is equivalent to the component-wise convergence.) For $n\geq 1$, let $\phi^n\colon [0,\,\infty)\to\mathbb{R}^d$ be right continuous functions having limits on the left on $(0,\,\infty)$, and let function $\phi\colon [0,\,\infty)\to\mathbb{R}^d$ be continuous. We say that ϕ^n converges to ϕ as $n\to\infty$ *uniformly on compacts* (u.o.c.) if for any $T\geq 0$,

$$||\phi^n - \phi||_T := \sup_{t\in[0,T]} |\phi^n(t) - \phi(t)| \to 0 \quad \text{as } n\to\infty\,,$$

and write it as $\lim\limits_{n\to\infty}\phi^n = \phi$. If function ϕ is differentiable at a point $s\in(0,\,\infty)$ then s is a *regular* point of ϕ, and we denote the derivative by $\dot\phi(s)$.

Recall that we consider a BC system with J infinite buffer servers and $K\geq J$ customer classes. In what follows, we use index k to denote the quantities related to class-k customers, $k\in\{1,2,\ldots,K\}$. Let $s(k)\subset\{1,\ldots,J\}$ be the set of servers that need to work together to service a class-k customer. Note that the capacity $\#s(k)\geq 1$ and that, if $\#s(k) = 1$, then server collaboration is not required. Evidently, $\cup_{k=1}^K s(k) = \{1,\ldots,J\}$, and we assume *non-overlapping* property: for each two classes $k\neq k'$,

$$s(k)\cap s(k') = \emptyset \quad \text{if} \quad \min\{s(k),\,s(k')\} > 1.$$

Define the customer classes $C(j) = \{k = 1,\ldots,K\,:\,j\in s(k)\}$ served by server $j\in\{1,\ldots,J\}$, and assume that, for each j, the capacity

$$\#\{k\in C(j)\,:\,\#s(k) > 1\}\leq 1.$$

In other words, at most one class may capture a given server for cooperation. To obtain *work-conserving* (or *non-idling*) discipline, we assume that multiserver customers have *preemptive-resume* priority.

Let $\xi_k(i)$, $i\geq 2$, be the independent identically distributed (i.i.d.) inter-arrival times of the ith class-k customers arriving from outside the system after instant 0, and let $\eta_k(i)$, $i\geq 2$, be the i.i.d. service times of the ith class-k customers finishing service after instant 0 (this is time required by any server in the set $s(k)$). All sequences are assumed to be mutually independent. We denote the generic elements of these sequences by ξ_k and η_k, respectively. The residual arrival time $\xi_k(1)$ of the first class-k customer entering the network after instant 0 is independent of $\{\xi_k(i),\,i\geq 2\}$. Also the residual service time $\eta_k(1)$ of a class-k customer initially being served, if any, is independent of $\{\eta_k(i),\,i\geq 2\}$, and $\eta_k(1) =_{st}\eta_k$ if class k is initially empty.

For each $k = 1,\ldots,K$, we impose the following standard conditions [3]:

$$\mathsf{E}\,\eta_k < \infty\,, \tag{1}$$

$$\mathsf{E}\,\xi_k < \infty\,, \tag{2}$$

$$\mathsf{P}(\xi_k\geq x) > 0\,, \quad \text{for any } x\in[0,\,\infty)\,. \tag{3}$$

Then, in particular, the arrival rate $\alpha_k := 1/\mathsf{E}\xi_k \in (0, \infty)$ and the service rate $\mu_k := 1/\mathsf{E}\eta_k > 0$, and we denote $\alpha = (\alpha_1, \ldots, \alpha_K)^T$ and $\mu = (\mu_1, \ldots, \mu_K)^T$. Also we assume that the inter-arrival times are *spread out*, that is, for some integer $r > 1$ and functions $f_k \geq 0$ with $\int_0^\infty f_k(y)\, dy > 0$,

$$\mathsf{P}\left(a \leq \sum_{i=2}^{r} \xi_k(i) \leq b\right) \geq \int_a^b f_k(y)\, dy, \quad \text{for any } 0 \leq a < b\,. \tag{4}$$

A class-k customer, when finishes service, re-enters the system and becomes class-ℓ customer with a probability $P_{k\ell} \in [0, 1)$. Then, with probability $1 - \sum_{\ell=1}^K P_{k\ell} \geq 0$, class-$k$-customer leaves the system upon service. Thus, $P := (P_{k\ell})_{k,\ell=1}^K$ is the (sub-stochastic) *routing* (or flow) matrix of the network. It is assumed that spectral radius of P is strictly less than 1, and hence, the inverse matrix $Q = (I - P^T)^{-1}$ is well defined. Define vector $\lambda = (\lambda_1, \ldots, \lambda_K)^T$ as (the unique) solution to the traffic equation

$$\lambda = \alpha + P^T \lambda, \quad \text{equivalently, } \lambda = Q\,\alpha,$$

where λ_k can be interpreted as the potential long run arrival rate of class-k customers into the system. Let $\rho_j = \sum_{k \in C(j)} \lambda_k/\mu_k$ be the traffic intensity for server j, and $\rho := (\rho_1, \ldots, \rho_J)^T$.

Now we introduce the following primitive processes describing the dynamics of the queueing network:

the exogenous arrival process $E = \{E(t) := (E_1(t), \ldots, E_K(t))^T, t \geq 0\}$, where

$$E_k(t) = \max\{\, n \geq 1 \,:\, \sum_{i=1}^{n} \xi_k(i) \leq t\}$$

is the total number of class-k arrivals from outside to the system in interval $[0, t]$. We also introduce the process $S = \{S(t) := (S_1(t), \ldots, S_K(t)), t \geq 0\}$, where the renewal process

$$S_k(t) = \max\{\, n \geq 1 \,:\, \sum_{i=1}^{n} \eta_k(i) \leq t\}$$

is the total number of class-k customers that would be served in interval $[0, t]$, provided all servers from $s(k)$ devote all time to class-k customers. (By definition, $E(0) = S(0) = 0$.) The routing process $\Phi = \{\Phi(n)\}_{n \in \mathbb{N}}$ is defined as follows:

$$\Phi_k(n) = \sum_{i=1}^{n} \phi^k(i)\,,$$

where, for each $i \in \mathbb{N}$, K-dimensional vectors $\phi^k(i) = \{\phi^k_\ell(i),\, \ell = 1, \ldots, K\}$ are i.i.d. (independent of the inter-arrival and service time processes), with at most one component equals 1, and the rest components being equal 0. If $\phi^k_j(i) = 1$ then

the ith class-k customer becomes class-j, while $\phi^k(i) = 0$ means the departure from the network.

Now we introduce the descriptive processes to measure the performance of the network. For any $t \geq 0$ and k, let $A_k(t)$ be the number of class-k arrivals (from outside and by feedback) by time t, $D_k(t)$ be the number of class-k departures (to other classes or outside the system), and let $Z_k(t)$ be the number of class-k customers being served at time t, so $Z_k(t) \in \{0,\, 1\}$. Also let $T_k(t)$ be the total service time devoted to class-k customers in interval $[0,\, t]$. Denote $Y_j(t)$ the idle time of server j in $[0,\, t]$, and let $Q_j(t)$ be the number of customers in the buffer of station j at time t, $j \in \{1, \ldots, J\}$. In an evident notation, processes D, T and Y are non-decreasing and satisfy initial conditions $D(0) = T(0) = Y(0) = 0$. We note that $A(0) = 0$, and assume that $Z(0)$ and $Q(0)$ are mutually independent and independent of all above given quantities.

For each t and k, we define the remaining time $U_k(t)$ until the next exogenous class-k arrival, and the remaining service time $V_k(t)$ of class-k customer being served at time t, if any. We introduce (in an evident notation) processes U and V, assume that they are right-continuous, and define $V_k(t) = 0$ if $Z_k(t) = 0$. Note that $U_k(0) = \xi_k(1)$, while $V_k(0) = \eta_k(1)$ if $Z_k(0) = 1$. Now we define the process $X = \{X(t),\, t \geq 0\}$ describing the dynamics of the network, where $X(t) := (Q(t), Z(t), U(t), V(t))^{\mathbf{T}}$, with the state space $\mathbb{X} = \mathbb{Z}_+^K \times \{0,1\}^K \times \mathbb{R}_+^K \times \mathbb{R}_+^K$. The process X is a piecewise-deterministic Markov process which satisfies Assumption 3.1 [5], and is a strong Markov process (p. 58, [3]).

We define the *workload* process $W = \{W(t) := (W_1(t),\, \ldots,\, W_J(t))^T,\, t \geq 0\}$, where $W_j(t)$ is the (workload) time needed to complete service of all class-k customers present in the system at time t, for any $k \in C(j)$. We introduce the *cumulative service time* process

$$\Upsilon = \{\Upsilon(n) := (\Upsilon_1(n_1),\, \ldots,\, \Upsilon_K(n_K))^T,\, n = (n_1, \ldots, n_K) \in \mathbb{N}^K\},$$

where $\Upsilon_k(n_k)$ is the total amount of service time of the first n_k class-k customers (including the remaining service time at time 0 for the first one), by any of the servers in the set $s(k)$. Note that this time is the same for each server from $s(k)$, and that $\Upsilon_k(0) = 0$.

The following *queueing network equations*, which are easy to verify, hold for all $t \geq 0$, $k = 1, \ldots, K$ and $j = 1, \ldots, J$:

$$A(t) = E(t) + \sum_{k=1}^{K} \Phi_k(D_k(t))\,, \tag{5}$$

$$D_k(t) = S_k\left(T_k(t)\right), \tag{6}$$

$$Q_k(t) = Q_k(0) + A_k(t) - \big(D_k(t) + Z_k(t)\big)\,, \tag{7}$$

$$\sum_{k \in C(j)} T_k(t) + Y_j(t) = t\,, \tag{8}$$

$$\int_0^\infty W_j(t)\, d\, Y_j(t) = 0\,, \tag{9}$$

$$W(t) = C\left(\Upsilon(Q(0) + A(t)) - T(t)\right), \tag{10}$$

where $e = (1, \ldots, 1)^T \in \mathbb{R}^d$ and C is the $J \times K$ matrix defined by:

$$C_{jk} = \begin{cases} 1, & \text{if } j \in s(k), \quad \text{equivalently, if and only if } \ k \in C(j), \\ 0, & \text{otherwise.} \end{cases}$$

Note that equation (9) reflects the work-conserving property introduced above. Also we note that equation (8) can be written as $C\,T(t) + Y(t) = t\,e$.

We assume that the service discipline is *head-of-the-line* (HL): only the oldest customer of each class can receive service. It gives the additional equation:

$$\Upsilon(D(t)) \leq T(t) < \Upsilon(D(t) + e)\,. \tag{11}$$

3 Stability Analysis of the BC system

By definition, a queueing network is *stable* if its associated underlying Markov process X is *positive Harris recurrent*, that is, it has a unique invariant probability measure. To prove stability of the network it is enough to establish stability of the associated fluid limit model [3].

3.1 The fluid limit model

Now we present, without proof, an analogue of Theorem 4.1 [3] (see also Proposition 1 [7]). If $X(0) = (Q(0), Z(0), U(0), V(0))^{\mathbf{T}} = x$, then we denote X as X^x (and analogously, for the processes E, S, D, T, Y, W).

Proposition 1. *Consider the BC system. Then, for almost all sample paths and any sequence of initial states $\{x_n\}_{n\geq 1} \subset \mathbb{X}$ with $\lim_{n\to\infty} |x_n| = \infty$, there exists a subsequence $\{x_{n_r}\}_{r\geq 1} \subseteq \{x_n\}_{n\geq 1}$ with $\lim_{r\to\infty} |x_{n_r}| = \infty$ such that the following limit*

$$\lim_{r\to\infty} \frac{1}{|x_{n_r}|}\, X^{x_{n_r}}(0) := \bar{X}(0)\,, \tag{12}$$

exists, and moreover the following u.o.c. limit exists for each $t \geq 0$,

$$\lim_{r\to\infty} \frac{1}{|x_{n_r}|} \Big(X^{x_{n_r}}(|x_{n_r}|t),\, D^{x_{n_r}}(|x_{n_r}|t),\, T^{x_{n_r}}(|x_{n_r}|t),\, Y^{x_{n_r}}(|x_{n_r}|t),\, W^{x_{n_r}}(|x_{n_r}|t) \Big)$$
$$:= \Big(\bar{X}(t),\, \bar{D}(t),\, \bar{T}(t),\, \bar{Y}(t),\, \bar{W}(t) \Big)\,, \tag{13}$$

where (in evident notation)

$$\bar{X}(t) := \Big(\bar{Q}(t),\, \bar{Z}(t),\, \bar{U}(t),\, \bar{V}(t) \Big)^{\mathbf{T}}\,,$$

and the components of vectors $\bar{U}(t)$, $\bar{V}(t)$ have, respectively, the form

$$\bar{U}_k(t) = (\bar{U}_k(0) - t)^+\,, \quad \bar{V}_k(t) = (\bar{V}_k(0) - t)^+,\ k = 1, \ldots, K\,. \tag{14}$$

Furthermore, the following equations are satisfied for any $t \geq 0$, $k = 1, \ldots, K$ *and* $j = 1, \ldots, J$*:*

$$\bar{A}(t) = t\,\alpha + P^T\,\bar{D}(t)\,, \tag{15}$$

$$\bar{D}(t) = M^{-1}\,\bar{T}(t)\,, \tag{16}$$

$$\bar{Z}_k(t) = 0\,, \tag{17}$$

$$\bar{Q}(t) = \bar{Q}(0) + \bar{A}(t) - \bar{D}(t) = \bar{Q}(0) + t\,\alpha - (I - P^T)\,\bar{D}(t)\,, \tag{18}$$

$$C\,\bar{T}(t) + \bar{Y}(t) = t\,e\,, \tag{19}$$

$$\int_0^\infty \bar{W}_j(t)\,d\,\bar{Y}_j(t) = 0\,, \tag{20}$$

$$\bar{W}(t) = C\,\big(M\,(\bar{Q}(0) + \bar{A}(t)) - \bar{T}(t)\big) = C\,M\,\bar{Q}(t)\,, \tag{21}$$

where diagonal matrix M *is defined as*

$$M = diag\big((\frac{1}{\mu_1}, \ldots, \frac{1}{\mu_k})^T\big).$$

We note that

$$\rho = C\,M\,\lambda. \tag{22}$$

Any limit $(\bar{X},\, \bar{D},\, \bar{T},\, \bar{Y},\, \bar{W})$ in (12), (13) is called a *fluid limit* associated with the BC system, [3]. Thus, Proposition 1 states that any *fluid limit* associated with the BC system satisfies the *fluid model equations* (15)-(21).

Remark 1. By Lemma 5.3 in [3], hereinafter we will assume without loss of generality that $\bar{U}(0) = \bar{V}(0) = 0$, which, by (14), implies $\bar{U}(t) = \bar{V}(t) = 0$ for all $t > 0$. We denote it $\bar{U} = \bar{V} = 0$ and identify $\bar{X}$ with $\bar{Q}$.

Definition 1. *The fluid limit* $(\bar{Q}, \bar{D}, \bar{T}, \bar{Y},\, \bar{W})$ *associated with a queueing network is stable, if there exists* $t_1 \geq 0$ *(depending on the input and service rates only) such that if* $|\bar{Q}(0)| = 1$*, then*

$$\bar{Q}(t) = 0 \;\; \textit{for all}\; t \geq t_1\,. \tag{23}$$

3.2 The stability criterion

Now we are ready to introduce and prove the stability criterion of the BC system, following Theorem 5.1 [3]. As in [6], the crucial fact in the proof is that the fluid limit $\bar{W}$ turns out to be a part of a solution of a *linear Skorokhod problem*, while the fluid limit process $\bar{Q}$ is instead used in [3]. We note that in some settings, the workload is better adapted to the use of the methodology of the Skorokhod problems than the queue-size process. On the other hand, a key point in the proof is the adequate choice of the *Lyapunov function*.

We prove the stability criterion under a technical Assumption (A) concerning the routing matrix P. We first introduce the process $\widetilde{W}$ as

$$\widetilde{W}(t) = C\,M\,Q\,\bar{Q}(t), \quad t \geq 0\,.$$

Assumption (A): The matrix P is such that for any $t \geq 0$,

$$\bar{W}_j(t) = 0 \text{ if and only if } \widetilde{W}_j(t) = 0,\ j = 1, \ldots, J.$$

Remark 2. Assumption (A) is trivially accomplished if no feedback is allowed since in this case $P \equiv 0$ and $Q = I$, implying $\widetilde{W} = \bar{W}$. For a non-trivial example of W-model with $J = 2$ and $K = 3$ (see Introduction), we easily find that

$$P = \begin{pmatrix} p_{11} & 0 & 0 \\ 0 & p_{22} & 0 \\ p_{31} & p_{32} & p_{33} \end{pmatrix}.$$

In other words, a feedback is allowed from class-j customers to class-j customers ($j = 1, 2$), and from class-3 customers to any class. Since $C = \begin{pmatrix} 1\,0\,1 \\ 0\,1\,1 \end{pmatrix}$ for the W-model, then we obtain

$$\widetilde{W}_1(t) = \frac{1}{\mu_1 (1 - p_{11})} \bar{Q}_1(t) + \big(\frac{p_{31}}{\mu_1 (1 - p_{11}) (1 - p_{33})} + \frac{1}{\mu_3 (1 - p_{33})}\big) \bar{Q}_3(t)\,,$$
$$\widetilde{W}_2(t) = \frac{1}{\mu_2 (1 - p_{22})} \bar{Q}_2(t) + \big(\frac{p_{32}}{\mu_2 (1 - p_{22}) (1 - p_{33})} + \frac{1}{\mu_3 (1 - p_{33})}\big) \bar{Q}_3(t).$$

Because

$$\bar{W}_1(t) = \frac{1}{\mu_1} \bar{Q}_1(t) + \frac{1}{\mu_3} \bar{Q}_3(t)\,,$$
$$\bar{W}_2(t) = \frac{1}{\mu_2} \bar{Q}_2(t) + \frac{1}{\mu_3} \bar{Q}_3(t)\,,$$

and all coefficients are positive, then Assumption (A) holds.

Theorem 1. *If the BC system with feedback given by fluid model equations* (15)-(21)*, satisfy conditions* (1)-(4) *and Assumption (A), then sufficient stability condition is*

$$\max_{j=1,\ldots,J} \rho_j < 1, \tag{24}$$

while $\max_{j=1,\ldots,J} \rho_j \leq 1$ *is the necessary stability condition.*

Proof. Sufficiency: By the equations (15)-(21),

$$\bar{A}(t) = t\,\alpha + P^T \left(\bar{Q}(0) + \bar{A}(t) - \bar{Q}(t)\right),$$

implying

$$\left(I - P^T\right) \bar{A}(t) = t\,\alpha + P^T \left(\bar{Q}(0) - \bar{Q}(t)\right).$$

It in turn implies

$$\bar{A}(t) = t\,\lambda + Q\,P^T \left(\bar{Q}(0) - \bar{Q}(t)\right). \tag{25}$$

By (25) we obtain

$$\begin{aligned}\bar{W}(t) &= C\,M\,\bar{Q}(0) + C\,M\,\bar{A}(t) - C\,\bar{T}(t)\\ &= C\,M\,\bar{Q}(0) + C\,M\left(t\,\lambda + Q\,P^T\left(\bar{Q}(0) - \bar{Q}(t)\right)\right) - t\,e + \bar{Y}(t)\\ &= C\,M\,\bar{Q}(0) + (\rho - e)\,t + C\,M\,Q\,P^T\left(\bar{Q}(0) - \bar{Q}(t)\right) + \bar{Y}(t)\\ &= C\,M\,Q\,\bar{Q}(0) + (\rho - e)\,t - C\,M\,Q\,P^T\,\bar{Q}(t) + \bar{Y}(t)\,. \end{aligned} \tag{26}$$

Since $\bar{W}(t) = C\,M\,\bar{Q}(t)$, it then follows from (26) that

$$C\,M\,Q\,\bar{Q}(t) = C\,M\,Q\,\bar{Q}(0) + (\rho - e)\,t + \bar{Y}(t)\,,$$

or, denoting $\widetilde{X}(t) = C\,M\,Q\,\bar{Q}(0) + (\rho - e)\,t$,

$$\widetilde{W}(t) = \widetilde{X}(t) + \bar{Y}(t).$$

It is easy to check that the following properties hold:

a) $\widetilde{X}(\cdot)$ has continuous paths with $\widetilde{X}(0) \geq 0$,
b) $\widetilde{W}(t) \geq 0$ for all $t \geq 0$,
c) $\bar{Y}(\cdot)$ has nondecreasing paths, $\bar{Y}(0) = 0$, and $\bar{Y}_j(\cdot)$ increases only at times t such that $\bar{W}_j(t) = 0$ for $j = 1, 2$ (see (20)). By Assumption (A), $\bar{Y}_j(\cdot)$ increases only when $\widetilde{W}_j(t) = 0$, $j = 1, 2$.

It follows that the paths of processes $(\widetilde{W}, \bar{Y})$ are solutions of the *continuous dynamic complementarity problem* (DCP) for $\widetilde{X}$ (see Definition 5.1 [3]), also known as the *deterministic Skorokhod problem*. Moreover, it is easy to check that condition (5.1) in [3],

$$\widetilde{W}(s) + \widetilde{X}(t+s) - \widetilde{X}(s) \geq \theta\,t \quad \forall t,\, s \geq 0\,,$$

is satisfied with $\theta := \rho - e$. Therefore, by Lemma 5.1 [3],

$$\dot{\bar{Y}}(s) \leq (e - \rho), \text{ if } s \geq 0 \text{ is a regular point of } \bar{Y}(\cdot). \tag{27}$$

Define function f as

$$f(t) = |\widetilde{W}(t)| = e^T\,\widetilde{W}(t).$$

It follows that

$$\begin{aligned} f(t) &= e^T\,(\widetilde{X}(t) + \bar{Y}(t)) = f(0) + e^T\,((\rho - e)\,t + \bar{Y}(t))\\ &= f(0) + \sum_{j=1}^{J}\left((\rho_j - 1)\,t + \bar{Y}_j(t)\right). \end{aligned} \tag{28}$$

Assume that $t > 0$ is a regular point for $\widetilde{W}$ (equivalently, for $\bar{Y}$).

If $f(t) > 0$, then there exists $j_0 \in \{1, \ldots, J\}$ such that $\widetilde{W}_{j_0}(t) > 0$, implying $\dot{\bar{Y}}_{j_0}(t) = 0$, by Assumption (A) and (20). Hence, by (28) and (27),

$$\begin{aligned}\dot{f}(t) &= \sum_{j=1}^{J} \big((\rho_j - 1) + \dot{\bar{Y}}_j(t)\big) = (\rho_{j_0} - 1) + \sum_{j \neq j_0} \big((\rho_j - 1) + \dot{\bar{Y}}_j(t)\big) \\ &\leq \rho_{j_0} - 1 \leq \max_{j=1,\ldots,J} \rho_j - 1 = -\kappa,\end{aligned}$$

where $\kappa = 1 - \max_{j=1,\ldots,J} \rho_j > 0$ by assumption. As f is a nonnegative function that is absolutely continuous and, for almost surely all regular points t, $\dot{f}(t) \leq -\kappa$ whenever $f(t) > 0$, then, by Lemma 5.2 [3], f is non increasing and $f(t) = 0$ for $t \geq f(0)/\kappa$. That is,

$$\widetilde{W}(t) = 0, \;\; t \geq \delta := \frac{|\bar{W}(0)|}{1 - \max_{j=1,\ldots,J} \rho_j}.$$

Finally, by Assumption (A) and (21), $\widetilde{W} = 0$ if and only if $\bar{Q} = 0$. Moreover, since

$$|\widetilde{W}(t)| = |C\, M\, Q\, \bar{Q}(t)| = \sum_{j=1}^{J} \Big(\sum_{k \in C(j)} a_{kj} \bar{Q}_k(t) \Big),$$

where a_{kj} depends on μ and matrix $Q = (q_{k\ell})_{k,\ell=1,\ldots,K}$. More exactly, $0 \leq a_{kj} \leq M_K$, for any $k = 1, \ldots, K$ and $j = 1, \ldots, J$, where

$$M_K = \big(\max_{k=1,\ldots,K} \max_{\ell=1,\ldots,K} q_{k\ell} \big) \big(\max_{k=1,\ldots,K} \frac{1}{\mu_k} \big)\, K > 0.$$

Then,

$$|\widetilde{W}(t)| \leq J\, M_K\, |\bar{Q}(t)|,$$

and we obtain

$$\bar{Q}(t) = 0, \quad t \geq \frac{J\, M_K\, |\bar{Q}(0)|}{1 - \max_{j=1\ldots,J} \rho_j} \geq 0.$$

It means that the fluid model is stable (by Definition 1), and Theorem 4.2 [3] ensures the stability of the queueing network.

Necessity: To prove the necessity of condition $\max_{j=1,\ldots,J} \rho_j \leq 1$, we assume $\rho_{j_0} > 1$ for some $j_0 \in \{1, \ldots, J\}$. Consider the non-negative function

$$g(t) = \widetilde{W}_{j_0}(t) = g(0) + (\rho_{j_0} - 1)\, t + \bar{Y}_{j_0}(t) \geq (\rho_{j_0} - 1)\, t > 0, \quad t > 0.$$

Then $\widetilde{W}_{j_0}(t) > 0$, which is equivalent to $\bar{W}_{j_0}(t) > 0$ by Assumption (A). By (21),

$$\bar{W}_{j_0}(t) = \sum_{k \in C(j_0)} \frac{1}{\mu_k}\, \bar{Q}_k(t)\,,$$

and hence $\bar{Q}(t) \neq 0$, finishing the proof. $\square$

We note that in practice, condition (24) can be treated as stability criterion which, for the W-model in Remark 2, becomes

$$\max\{\rho_1,\, \rho_2\} < 1,$$

where

$$\rho_1 = \alpha_1 \frac{1}{\mu_1\,(1-p_{11})} + \alpha_3 \left(\frac{p_{31}}{\mu_1\,(1-p_{11})\,(1-p_{33})} + \frac{1}{\mu_3\,(1-p_{33})} \right)$$
$$\rho_2 = \alpha_2 \frac{1}{\mu_2\,(1-p_{22})} + \alpha_3 \left(\frac{p_{32}}{\mu_2\,(1-p_{22})\,(1-p_{33})} + \frac{1}{\mu_3\,(1-p_{33})} \right).$$

This can be easily seen since $\rho = C\,M\,Q\,\alpha$,

$$C\,M = \begin{pmatrix} \frac{1}{\mu_1} & 0 & \frac{1}{\mu_3} \\ 0 & \frac{1}{\mu_2} & \frac{1}{\mu_3} \end{pmatrix}$$

and

$$Q = (I - P^T)^{-1} = \begin{pmatrix} \frac{1}{1-p_{11}} & 0 & \frac{p_{31}}{(1-p_{11})\,(1-p_{33})} \\ 0 & \frac{1}{1-p_{22}} & \frac{p_{32}}{(1-p_{22})\,(1-p_{33})} \\ 0 & 0 & \frac{1}{1-p_{33}} \end{pmatrix}.$$

If the model does not allow feedback, then $p_{ij} = 0$ for all $i, j = 1, 2, 3$, and

$$\rho_1 = \frac{\alpha_1}{\mu_1} + \frac{\alpha_3}{\mu_3}, \quad \rho_2 = \frac{\alpha_2}{\mu_2} + \frac{\alpha_3}{\mu_3}.$$

4 Conclusion

We consider a *Basic Collaboration* queueing system, which is a multiclass queueing system with feedback, that generalizes the so-called W-model [14]. In the system, some customer classes cooperate to be served by a subset of non-overlapping servers. We apply the fluid limit approach methodology [3] to find stability condition of the system.

References

1. Arthurs, E., Kaufman, J.S.: Sizing a message store subject to blocking criteria. In: Proceedings of the Third International Symposium on Modelling and Performance Evaluation of Computer Systems: Performance of Computer Systems, pp. 547–564. North-Holland Publishing Co., Amsterdam, The Netherlands (1979)
2. Brill, P., Green, L.: Queues in which customers receive simultaneous service from a random number of servers: A system point approach. Management Science. 30(1), 51–68 (1984)

3. Dai, J.G.: On positive Harris recurrence of multiclass queueing networks: a unified approach via fluid limit models. Ann. Appl. Prob. 5(1), 49–77 (1995)
4. Dai J.G.: A fluid limit model criterion for unstability of multiclass queueing networks. Ann. Appl. Prob. 6(3), 751–757 (1996)
5. Davis, M.H.A.: Piecewise deterministic Markov processes: a general class of non-diffusion stochastic models. J. Roy. Statist. Soc. Ser. B. 46, 353–388 (1984)
6. Delgado, R.: State space collapse and stability of queueing networks. Math Meth Oper Res 72, 477-499 (2010)
7. Delgado, R., Morozov, E.: Stability analysis of cascade networks via fluid models. Performance Evalutation 82, 39-54 (2014)
8. Fletcher, G.Y., Perros, H., Stewart, W.: A queueing system where customers require a random number of servers simultaneously. European Journal of Operational Research 23, 331–342 (1986)
9. Garnet, O., Mandelbaum, A.: An introduction to Skills-Based Routing and its operational complexities, http://iew3.technion.ac.il/serveng/Lectures/SBR.pdf
10. Green, L.: Comparing operating characteristics of queues in which customers require a random number of servers. Management Science 27(1), 65–74 (1980)
11. Kaufman, J.: Blocking in a shared resource environment. IEEE Transactions on Communications 29(10), 1474–1481 (1981)
12. Kim, S.: M/M/s queueing system where customers demand multiple server use. PhD thesis, Southern Methodist University (1979)
13. Rumyantsev, A., Morozov, E.: Stability criterion of a multiserver model with simultaneous service. Ann Oper Res (2015). doi:10.1007/s10479-015-1917-2
14. Talreja, R., Whitt, W.: Fluid models for overloaded multiclass many-server queueing systems with first-come, first-served routing. Management Science 54, 1513-1527 (2008)
15. Tikhonenko, O.: Generalized Erlang problem for service systems with finite total capacity. Problems of Information Transmission 41(3), 243-253 (2005)
16. Van Dijk, N.M.: Blocking of finite source inputs which require simultaneous servers with general think and holding times. Oper Res Lett 8(1), 45-52 (1989)
17. Whitt, W.: Blocking when service is required from several facilities simultaneously. AT&T Technical Journal 64(8), 1807-1856 (1985)
18. Wong, D., Paciorek, N., Walsh, T., DiCelie, J., Young, M., Peet, B.: Concordia: An infrastructure for collaborating mobile agents. International Workshop on Mobile Agents MA 1997: Mobile Agents. LNCS, vol. 1219, pp. 86-97. Springer (1997)